Grading Degradation in an Institutionally Managed Repository

Luis Meneses, Sampath Jayarathna, Richard Furuta and Frank Shipman

Center for the Study of Digital Libraries and Department of Computer Science and Engineering
Texas A&M University, College Station, TX 77843-3112 USA
(ldmm, sampath, furuta, shipman)@cse.tamu.edu

Imagine a library filled with books that have missing pages. It might seem as overly exaggerated, but that metaphor can be used to depict the state of the digital repositories that have been affected by unexpected change. Many research groups (including ours) have examined the implications of the Web's decentralized administration on the stability of materials. [1-4]. We have found that electronic resources can change, both intentionally and unintentionally, because of different factors and circumstances. Change can occur because of deliberate actions on the part of the collector, unexpected events or may be due to other uncontrollable factors.

A natural assumption we make is that some sources are highly curated – hence accurate – and others are less institutional and hence more likely to exhibit symptoms of change. However, this is not often the case. For example, we observed that the ACM Digital Library had 15 unique links referencing the different sites in the JCDL conference series and 8 of them report errors or point to the wrong content. Therefore, we have been examining a case study that shows that this assumption does not always hold and also helps illustrate some of the complexities on change in the current Web environment.

The corpus for our case study is the Association for Computing Machinery list of conference proceedings (http://dl.acm.org/proceedings.cfm), which we retrieved on 9/27/2014. We were able to extract 6086 conference URLs – out of which 2001 were unique. We then categorized the retrieved pages according to their HTTP response codes and categorized the 1492 pages with a 200 HTTP response code. As a result of this categorization, we found that 917 pages were "clearly correct", 531 were incorrect and 44 didn't provide us enough information to make an accurate assessment. The 531 pages that were reported by the HTTP server as being correctly retrieved but were clearly not the original contents were then analyzed in an effort to understand how conference sites degrade over time. Figure 1 shows the distribution of the incorrect pages.

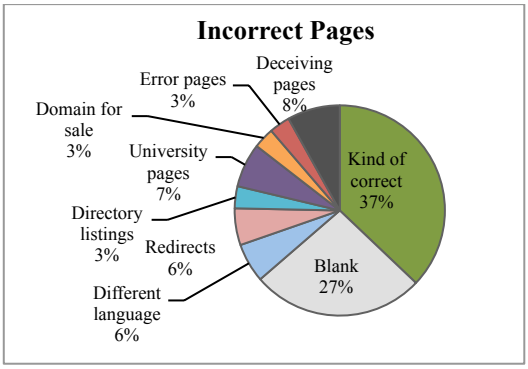

Figure 1: Distribution of the incorrect pages.

In the end, we were able to develop a classification system that identifies nine categories that we used to group the "incorrect" pages. These nine groups (which are shown in Figure 1) provide insight regarding the different stages that conference pages go through until they are ultimately abandoned, and help illustrate some of the complexities of change in the Web. On the other hand, we found the documents in the "Deceiving pages" category particularly interesting. These are pages that have been taken over by a third party and the content displayed in them is totally unrelated to the original purpose of the site.

We also found that many of these pages have a suspicious purpose: they were not created to deceive users, but as an attempt to manipulate the PageRank algorithm [5]. Most notably, the documents in the "deceiving pages" category share two characteristics. First, they had a greater number of links pointing to other pages within the site when compared to the number of out-links; and second, the domain names that host these pages once belonged to a reputable institution for number of years (i.e., a conference series) before being abandoned. Consequently, these abandoned domain names have value – not necessarily due to current network traffic but in the perception of their authority/validity. This problem becomes increasingly interesting when we consider that the cost of creating a web page is small and that some search engines (most notably Google) do not share the overall rankings for their indexed sites, which can lead some parties to abuse these malicious techniques.

REFERENCES

[1] M. Klein, J. Ware, and M. L. Nelson, "Rediscovering missing web pages using link neighborhood lexical signatures," in *Proceedings of the 11th annual international ACM/IEEE Joint Conference on Digital libraries*, Ottawa, Ontario, Canada, 2011.

[2] H. M. SalahEldeen and M. L. Nelson, "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?," in *Proceedings of Theory and Practice of Digital Libraries 2012*, Paphos, Cyprus, 2012.

[3] L. Francisco-Revilla, F. Shipman, R. Furuta, U. Karadkar, and A. Arora, "Managing change on the web," in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia, United States, 2001.

[4] L. Meneses, H. Barthwal, S. Singh, R. Furuta, and F. Shipman, "Restoring Semantically Incomplete Document Collections Using Lexical Signatures," in *Proceedings of Theory and Practice of Digital Libraries 2013*, Valletta, Malta, 2013.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford University, 1999, http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.